AC

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
19 July 2001 (19.07.2001)

PCT

(10) International Publication Number
WO 01/51518 A2

(51) International Patent Classification[7]: C07K 14/435

(21) International Application Number: PCT/US01/01275

(22) International Filing Date: 12 January 2001 (12.01.2001)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
09/483,618 14 January 2000 (14.01.2000) US

(71) Applicant *(for all designated States except US)*: **LUDWIG INSTITUTE FOR CANCER RESEARCH** [CH/US]; 605 Third Avenue, New York, NY 10158 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **SIMPSON, Andrew, John, George** [GB/BR]; Rua Prof. Antonio Prudente, 109-4° andar, CEP-01509-010 São Paulo, SP (BR). **CORREA, Ricardo** [BR/BR]; Rua Prof. Antonio Prudente, 109-4° andar, CEP-01509-010 São Paulo, SP (BR). **DE SOUZO, Sandro** [BR/BR]; Rua Prof. Antonio Prudente, 109-4° andar, CEP-01509-010 São Paulo, SP (BR).

(74) Agent: **HANSON, Norman, D.**; Fulbright & Jaworski L.L.P., 666 Fifth Avenue, New York, NY 10103 (US).

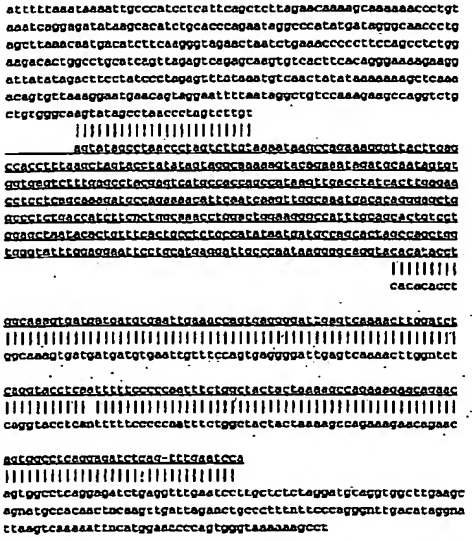
(81) Designated States *(national)*: AU, CA, CN, JP, KR, US.

(84) Designated States *(regional)*: European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR).

Published:
— *without international search report and to be republished upon receipt of that report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

(54) Title: ISOLATED NUCLEIC ACID MOLECULE ENCODING A HUMAN SEMAPHORIN MOLECULE, AND USES THEREOF



(57) Abstract: The invention relates to isolated nucleic acid molecules which encode human semaphorin R/6B. These molecules are associated with cancer, such as breast cancer. Also a part of the invention are the proteins encoded by these molecules, as well as oligonucleotides which hybridize to them. Various uses of these molecules are disclosed as well.

WO 01/51518 A2

ISOLATED NUCLEIC ACID MOLECULE ENCODING A HUMAN SEMAPHORIN MOLECULE, AND USES THEREOF

RELATED APPLICATIONS

This application is a continuation in part of Serial No. 09/483,618, filed January 14, 2000, which is a continuation-in-part of Serial No. 09/406,117, filed September 27, 1999, which is a continuation in part of Serial No. Application Serial Number 09/196,716, filed on November 20, 1998, the disclosure of which is incorporated by reference in its entirety. It is also a continuation-in-part of PCT application PCT/US99/27430, filed November 19, 1999, designating the United States.

FIELD OF THE INVENTION

The invention relates to isolated nucleic acid molecules which encode human analogs of semaphorin, the proteins encoded thereby, as well as their use. The molecules described herein were isolated and identified using the ORESTES method, which is described herein.

BACKGROUND AND PRIOR ART

The area of nucleic acid research has seen tremendous advances in knowledge and understanding in the recent past. One of the goals in the field has been the determination of the sequence of the entire chromosomal component, or "genome" of organisms. This has been achieved for several non-nucleated organisms (prokaryotes), and of one organism with a nucleus, a "eukaryote". Eukaryotes have much more complex genomes than prokaryotes, for reasons which will be discussed infra.

The interest in sequencing entire genomes of organisms has been explained in detail in both technical and non-technical publications, and need not be repeated here. See, for example Venter, et al, "Shotgun Sequencing of The Human Genome", Science 280:1640 - 1642 (1998), Pennisi, "A Planned Boost for Genome Sequencing, But the Plan Is in Flux", Science 281: 148-149 (1998).

Various approaches to what is a large, and complex project have been advanced. For example, the so-called "Shotgun" approach, developed by Venter et al, is very well known. In this approach, genomic DNA is cleaved into very small pieces, and these pieces are then sequenced. The approach is repeated, and after an undefined number of repeats, sequences are aligned to permit, at least in theory, a determination of the complete genomic sequence.

This approach has been used by Venter et al on prokaryotes, and it has been proposed for use on more complex eukaryotes, such as humans. The proposed approach to eukaryotes is not without drawbacks and criticism, however. A sizable portion of the scientific community is of the view that the resulting information will be riddled with gaps. The human genome, in contrast to prokaryotic genomes is characterized by a large number of repetitive sequences. It is felt by many that the overlapping of repetitive sequences could lead to incorrect alignment of the larger fragments from which they are derived.

A second approach, which has found more widespread acceptance, is to cleave the genome into relatively large fragments, and then to "map" the larger, non-sequenced fragments to show overlap prior to sequencing the material. After this overlapping, which results in a physical map of the genome, the segments are fragmented, and sequenced. While this approach should, in theory, eliminate the gaps in the sequence, it is time consuming and costly. Further, both of these approaches suffer from a fundamental drawback, as will all approaches which begin with eukaryotic genomic DNA, as will now be explained.

Eukaryotic DNA consists of both "coding" and "non-coding" DNA. For purposes of this invention, only coding DNA is under consideration, as it is this material which is transcribed and then translated into proteins. This coding DNA is sometimes referred to as "open reading frames" or "ORFs", and this terminology will be used hereafter.

As compared to prokaryotes, eukaryotic DNA has a much more complex structure. Genes generally consist of a non-coding, regulatory portion of hundreds of nucleotides followed by coding regions ("exons"), separated by non-coding regions ("introns"). When DNA is transcribed into messenger RNA, or mRNA, and then translated into protein, it is only these exons which are of interest. It has been estimated that, for humans, of the approximately 3 billion nucleotides which make up the genome, only about 3% are coding sequences. The shotgun and mapping approaches referred to supra do not differentiate

between coding and non-coding regions. Hence, a method which would permit sequencing of only coding regions would be of great interest, especially if the method permits development of longer "contigs" of sequence information.

One such method is, in fact known. This is the "Expressed Sequence Tag" or "EST" approach. In this approach, one works with complementary DNA or "cDNA" rather than genomic DNA. In brief, as indicated supra, genomic DNA is transcribed into mRNA. The mRNA contains the relevant ORF in contiguous form, i.e. without intervening introns. These molecules are very fragile and their existence transient. In the laboratory, one can employ various enzymes, i.e., so-called "reverse transcriptases" to prepare complementary DNA, or "cDNA", which is much more stable than mRNA. One then sequences the cDNA, incompletely, from either the 5' or 3' end. These incomplete sequences, in theory, serve as identifying "tags" for nucleic acid molecules of interest. Literally millions of ESTs have been prepared, and are accessible via known data bases, such as GenBank.

There are problems with this approach as well. First, large amounts of extremely high quality mRNA are necessary, and this is not always available. Also, one must bear in mind that the non-coding regions of mRNA molecules are found at the 5' and 3' ends, and this is carried over into the cDNA molecule. As a result, the information obtained may not be very useful. For example, it frequently provides no information about the actual protein encoded by the molecule. Clearly, there is a need for a system which provides more useful information about nucleic acid molecules.

Dias Neto et al., Gene 186: 135-142 (1997), the disclosure of which is incorporated by reference, applied a method for determining sequence information from the parasite S. mansoni which involved, inter alia, the use of arbitrary primers, and low stringency hybridization conditions. There is no discussion in this paper of the ability to identify and to sequence internal portions of an open reading frame. The paper itself appears to have only been cited a single time by other investigators. Nor is there any discussion within the reference of investigating sequences for overlap, so as to develop "contigs", i.e, longer nucleotide sequences prepared by determining overlap of two smaller sequences.

U.S. Patent No. 5,487,985 to McClelland, et al., incorporated by reference, teaches a method referred to as "AP-PCR", or arbitrarily primed polymerase chain reaction. The

method employs a single primer designed so that there is a degree of internal mismatch between the primer and the template. Following amplification with the primer, a second PCR is carried out. The amplification products are separated on a gel to yield a so-called "fingerprint" of the organism or individual under study. The '985 patent does not discuss the identification of internal portions of open reading frames, nor does it discuss the analysis of sequences to develop contigs.

The semaphorins are one of the most prominent of the conserved families of axon guidance molecules. As described by, e.g., Van Vactor et al., Curr Biol 25(9): R201-204(1999) they are expressed in many different regions of the developing nervous system, and are known to play important roles in establishing accurate axonal projections.

The semaphorins have been divided into 7 different classes, depending upon whether they are secreted molecules or transmembrane molecules. For example, the membranes of the Class III semaphorin family are secreted molecules, known to act as repulsive factors for specific axonal populations. "Semaphorin III," which has also been referred to as "Collapsin-I" or "Sema D," causes growth cone collapse, as well as axonal retraction and repulsion, in sensory and sympathetic axons in culture. See, e.g., Yu et al., Neuron22(1):11-14(1999).

Eckhardt, et al. Mol. Cell Neurosci9(5/6):409-19 (1997) have identified a murine semaphorin cDNA, which is referred to as "SemaVIb." The molecule comprises a characteristic, extracellular semaphorin domain, but lacks both the immunoglobin domain and thrombospondin repeats that have been observed in other vertebrate, transmembrane semaphorins. Sema VIb is expressed in subregions of the nervous system during development, and is especially prominent in muscle tissue. Sema VIb mRNA is expressed ubiquitously in adulthood. Studies carried out in vitro have shown that this molecule binds the SH3 domain of c-src. This may indicate a role in intracellular signaling via an src-related cascade.

Christensen, at al. Canc. Res 58(6): 1238-44 (1988) identified a murine semaphorin/collapsin family in metastatic murine mammary adenocarcinoma cell lines. The molecule was identified using differential display -PCR, and is now known as "M-Sema H." It is less abundant in normal tissues as compared to tumor cells. This work is significant because it is the first example of positive correlation of semaphorin expression with tumor progression,

and suggests a role for M-Sema H during metastasis and nerve axon development. Christensen et al. have deposited three sequences in GENBANK, referred to as Z93947, Z93948, and Z80941. Also see PCT Application W09947671 to Christensen et al., published on September 23, 1999.

There is an extensive patent literature on members of the semaphorin family. See, e.g., US Patent No. 5,981,222 to Jacobs, et al.; 5,935,865 to Goodman, et al.; 5,807,826 to Goodman, et al.; 5,639,856 to Goodman, et al. International applications, in addition to W09947671, described supra, include W09958676; W09945114; W09932622; W09921997; W09904263; W09902556; W09853065; W09822504; W09815628; W09811216; and W09720928. All of these U.S. and international applications, which represent some, but not all of the patent literature in this area, are incorporated by reference.

Of further interest is the work of Comoglio, et al., Exp Cell Res 253(1):88-99(1999), and Trusolino, et al, FASEB J 12(13):1267-80(1998). These references discuss so-called "scatter factors" and their receptors. Comoglio, et al, identify a new gene family of molecules which share homology to the scatter factor receptors, and identify these new molecules as semaphorin receptors, and suggest that deregulation of semaphorin may confer invasive and metastatic properties t cancer cells.

The ORESTES methodology, described herein, has been applied to human breast tumor cells, and a complete sequence of a human semaphorin related molecule has been identified. Homology analysis reveals that it is 84% identical to murine semaphorin VIb, discussed supra. This is the molecule with which it shares greatest identity. Different forms of the molecule have been identified as well.

Previously, a human molecule was identified and referred to as Semaphorin R. See, e.g., U.S. Patent application Serial No. 09/483,618, filed January 14, 2000, the disclosure of which is incorporated by reference. Recently, Bamberg et al., "Unified Nomenclature for the Semaphorins/Collapsins," Cell 97:551-552 (May 28, 1999), the disclosure of which is incorporated by reference, suggest renaming the members of the semaphorin family and, in accordance with this system, Semaphorin R would now be referred to as Semaphorin 6B.

How these molecules were identified, as well as their uses, will be clear from the disclosure which follows.

BRIEF DESCRIPTION OF THE FIGURES

Figures 1A and 1B both show, schematically, prior art genome sequencing approaches.

Figure 1C shows the invention, schematically.

Figure 2 presents both a theoretical probability curve (dark ovals) and actual results (white ovals), obtained when practicing the invention. The data points refer to the probability of securing the sequence of a particular portion of cDNA molecule when practicing the invention.

Figure 3 shows construction of a contig, using the invention.

DETAILED DESCRIPTION OF PREFERRED EMBODIMENTS

One aspect of the invention, as discussed supra, is a method for obtaining nucleotide sequence information from organisms, preferably information from open reading frames of cDNA of eukaryotic organisms. As a first step, messenger RNA ("mRNA") is extracted from a cell. The extraction of mRNA is a standard technique, the details of which are well known by the artisan of ordinary skill. For example, it is well known that eukaryotic mRNA, as compared to other forms of RNA, is characterized by a "poly A" tail. One can separate mRNA from other types of RNA by passing it over a column which contains oligomers of the base thymidine. These "oligo dT" molecules hybridize to the poly A sequences on the mRNA molecules, and these then remain on the column. Other approaches to separation of mRNA are known. All can be used. If prokaryotic mRNA is being considered, separation using poly A/poly T hybridization is not carried out. It is preferred to treat the resulting material to reduce or to eliminate contamination by DNA. Adding a DNA degrading enzyme, such as DNA ase is preferred. This is carried out prior to contact with the column. It is also preferred to pas the purified RNA over the column at least twice.

The separated mRNA is then used to prepare a cDNA. The preparation of the cDNA represents the first inventive step in the method of the invention. To prepare the cDNA, the mRNA is combined with a sample of a single, arbitrary primer. By "arbitrary" is meant that the primer used does not have to be designed to correspond to any particular mRNA molecule. Indeed, it should not be, because the primer is going to be used to make all of the

cDNA. Details on the design of arbitrary primers can be found in Dias-Neto, et al., supra, McClelland, et al., supra, and Serial No. 08/907,129 filed August 6, 1997 and incorporated by reference.

The primer is preferably at least 15 nucleotides long. Theoretically, it should not exceed about 50 nucleotides, but it can. Most preferably, the primer is 15-30 nucleotides long. While the sequence of the primer can be totally arbitrary, it is preferred that the total content of nucleotides "G" and "C" in the primer be compatible with the "G" and "C" content of the open reading frames of the organism under consideration. It is found that this favors amplification of the desired sequences. General rules of primer construction favor a G and C content of at least 50%.

"Arbitrary primer" as used herein does not exclude specific design choices within the primers. For example, the four bases at the 3' end of a given primer are generally considered the most important portion for hybridization. Hence, it is desirable to include as many different primers as possible, to cover all variations within this 4 base sequence. There are 256 variants possible, since there are four nucleotides. In order to identify products from a particular source, a "marker" sequence can be used, i.e., a stretch of predefined nucleotides. The remainder of the primer should be selected to correspond to overall GC usage, as described supra. Hence, for a primer 25 nucleotides long, the first 17 should correspond to GC usage for the organism in question. Nucleotides 18-21 would be a "tag", such as "GGCC." Then, all possible combinations of four nucleotides would follow, to produce 256 primers, which contain a known marker. This procedure could be repeated with a second set of primers, where the marker at 18-21 is different.

In practice, each set of variants is used with mRNA from a single source, and would permit the artisan to mark all sequences from a source, and still permit pooling.

The primer is combined with the mRNA under low stringency conditions. What is meant by this is that the conditions are selected so that the primer will hybridize to partially, rather than to only completely complementary sequences. Again, this is necessary because the primer will amplify an arbitrary sample of the mRNA pool, not just one sequence. There are standard rules and formulas for approximating high and low stringency, and the artisan of ordinary skill is familiar with these. Attention is drawn to Simpson, et al, U.S. Patent

application Serial No: 08/907,129, filed August 6, 1997, incorporated by reference, for more information on this, as well as Dias-Neto, et al. and McClelland, et al., supra.

The arbitrary primer and mRNA are mixed with appropriate reagents, such as reverse transcriptase, a buffer, and dNTPs, to yield a pool of single stranded, cDNA molecules.

Once the single stranded cDNA is prepared, it is used in an amplification reaction. In this second reaction, it is preferred, but not required, that the single primer used is identical to the first primer, as described supra, and that low stringency conditions be employed. Using identical primers tends to produce longer products, but this is not required.

The result of this amplification is a mini library. One can carry out cDNA synthesis in multiple, separate reactions, using different arbitrary primers, "A", "B", "C" and "D". Four pools of single stranded cDNA are then produced, i.e, "A", "B", "C" and "D". Each pool is then amplified using each of the four primers, to generate mini-libraries AA, AB, AC, AD, BA, BB, BC, BD, CA, CB, CC, CD, DA, DB, DC, and DD. These mini-libraries are used in the sequencing reaction which follows.

Once the cDNA is prepared, the resulting products are isolated, such as by size fractionation on a gel. The resulting bands can be removed from the gel, such as by elution, and then subjected to standard methodologies for cloning and sequencing.

Key to this feature of the invention, as is described herein, is the use of arbitrary primers under low stringency conditions. This combination permits the artisan to sequence internal regions of cDNA preferentially, as compared to the 5' and 3' ends, as is typical in standard prior art approaches. Specifically, consider a portion of a cDNA molecule which is a distance "S" from the 3' end of the molecule. For this portion of the molecule to be amplified by a primer, the primer must bind on both sides of the region to be amplified. If the complete length of the molecule is represented by "L", the probability of a primer binding to the nucleic acid molecule on both sides of a point on a nucleic acid molecule is S(L-S).

The highest probability for inclusion within amplified cDNA is the exact middle of the molecule. Lowest priority, in contrast, is at the extreme 5' and 3' ends. To elaborate, assume a point directly in the middle of a cDNA molecule, i.e., if the molecule is "x + 1" nucleotides long, .5x nucleotides precede the midpoint, and .5x nucleotides follow it. The likelihood of a primer hybridizing to a point on the molecule, preceding the middle is .5x, and following it is also .5x. If "x" is 1, then the probability of hybridization surrounding the midpoint is .5(1-.5), or .25, i.e., 25%. Similarly, assume a point on the same molecule located

.9x away from the 3' end. In this case, since the molecule is "x" units long, the point is .1x from the 5' end, i.e., .1 units precede it, and .9 units follow it. If the length is 1, then the probability of hybridization surrounding this process is .9 (1-.9), or 9%. Hence, by using a primer and conditions which permit hybridization of the primer anywhere along the molecule, one actually secures the majority of amplified products from within a cDNA molecule, rather than at the ends. In figure 2 of this application, one sees a curve which results when the theoretical model is applied (dark ovals), and a curve obtained in practice (light ovals). It will be seen that, remarkably, the practice of the invention is actually very close to the theory.

One very practical result of this approach is that the mRNA is normalized, and bias in copy number is eliminated. The probability of producing an EST from a given mRNA is proportional to the length of that molecule and not its abundance within the source being analyzed.

A further aspect of the invention is the construction of contigs, once the sequence information has been determined. One creates a contig by comparing sequence information and finding overlaps. For example, the last 300 nucleotides of a sequence may be identical to the first 300 nucleotides of a second sequence. The artisan can essentially splice the first and second sequences together, to produce a longer one. The splicing can be done with two or more sequences found in the particular experiment that is carried out, or by comparing deduced sequences to sequences which are available in a public data base, a private data base, a journal, or any other source of sequence information.

A further aspect of the invention is the ability to compare information obtained using the inventive method to pre-existing information, in order to determine if a known nucleotide sequence is an internal sequence of a particular gene. This can be done because, as explained supra, the method described herein generates an extremely high percentage of internal sequences, with a very low percentage of sequences at the ends of a given molecule. The prior art methods either generate predominantly terminal sequences, or internal sequences on a completely random basis. Hence, it is probable that nucleotide sequences of unknown origin are contained within various sources of sequence information. Data generated using the methods of this invention can be compared to this pre-existing information very easily, and can result in a determination that a particular nucleotide sequence is, in fact, an internal sequence.

The practice of the invention and how it is achieved will be seen in the examples which follow.

EXAMPLE 1

This example describes the generation of a cDNA library in accordance with the invention. While colon cancer cells from a human were used, any cell could also be treated in the manner described herein.

The mRNA was extracted from a sample of colon cancer cells, in accordance with standard methods well known to the artisan, and not repeated here. It was then divided into approximately 5$\mu\ell$ aliquots, which contained anywhere from 1 to 10ng of mRNA. The samples were then stored at -70°C until used.

The aliquots of mRNA were then used to prepare single stranded cDNA, using 25 pmol samples of a single, arbitrary primer. Several different experiments were carried out, using a different, single arbitrary primer in each case.

The single, arbitrary primers used were:

5' - GAAGCTGGTA	AACAAAAGG - 3'	SEQ ID NO: 8
5' - AGCTGCATGA	TGTGAGCAAG - 3'	SEQ ID NO: 9
5' - CCCGCTCCTC	CTGAGCACCC - 3'	SEQ ID NO: 10
5' - GAGTCGATTT	CAGGTTG - 3'	SEQ ID NO: 11
5' - TGCTTAAGTT	CAGCGGG - 3'	SEQ ID NO: 12

In each case, 25 pmols of arbitrary primer were mixed with the aliquot of mRNA, 100 units of Moloney murine leukemia virus reverse transcriptase, reverse transcriptase buffer (25mM Tris-HCl, pH 8.3, 75mm KCl, 3mM $MgCl_2$, 10mM DTT), and 100mM of each dNTP, to a final volume of 20uL. The mixture was incubated for 30 minutes, at 37°C, to yield single stranded cDNA.

EXAMPLE 2

The single stranded cDNA produced in example 1, supra, was used as the template in a PCR amplification reaction. In this, a sample of 1ul of single stranded cDNA was

combined, together with the same primer that had been used to generate the cDNA. Amplification was carried out, using 12uM of primer, 200 uM of each dNTP, 1.5mM $MgCl_2$, 1 unit of DNA polymerase, and buffer (50mM KCl, 10mM Tris-HCl, pH9.0, and 0.1% Triton X-100), to reach a final volume of 15ul. Then, 35 cycles of amplification were carried out, 1 cycle consisting of 95°C for 1 minute, (denaturation), 37°C for 1 minute (annealing), and extension at 72°C, for 1 minute. In the final cycle extension was increased for 5 minutes. The amplification products were used in the analyses which follow. Additional experiments were also carried out, in the same fashion, using different primers.

EXAMPLE 3

In order to analyze the amplification products, 3ul samples were mixed with 3ul of sample buffer, 0.05% bromophenol blue, 0.05% xylene cyanol FF, and 7% sucrose (w/v), in distilled water, and then visualized on silver stained, 6% polyacrylamide gels, following Sanguinetti, et al., Biotechniques 17:3-6 (1994), incorporated by reference.

The steps set forth supra result in banding patterns on the gel, each band representing a different sequence. The most complex banding patterns were analyzed, as discussed in example 4, infra. It is important to note that controls were run during the experiments, to make sure that genomic DNA had not contaminated the samples. In brief, the control experiments used mRNA and genomic DNA, without reverse transcription PCR. The profiles obtained should differ, in each case from those obtained using reverse transcribed mRNA, and did so.

EXAMPLE 4

The cDNAs generated in the preceding examples were mixed, by pooling 10-20ul of each set of products into a final volume of 60ul, followed by electrophoresis through a 1% low melting point agarose gel containing ethidium bromide to stain the cDNA fragments. Known DNA size standards were also provided.

The gel portions containing fragments between 0.25 and 1.5 kilobases were excised, using a sterile razor blade. Excised agarose was then heated to 65°C for 10 minutes, in 1/10 volume of NaOAc (3mM, pH 7.0), and cDNA was recovered via standard phenol/chloroform

extraction and ethanol precipitation, followed by resuspension in 40ul of water. The thus recovered cDNA was used in the following experiments.

EXAMPLE 5

The cDNA extracted supra was treated with 10 units of Klenow fragment cDNA polymerase, and 10 units of T4 polynucleotide kinase, for 45 minutes at 37°C. The reaction mixture was then extracted, once, with phenol, and the DNA was then recovered by passage through a standard Sephacryl S-200 column. Recovered cDNA was then ligated into the commercially available plasmid pUC18, and the plasmids were used to transform receptive E. coli, using standard methodologies. This resulted in sufficient amounts of individual cDNA molecules for the experiments which follow.

EXAMPLE 6

Individual bacterial clones were established from the transformants of example 5. These were then used to prepare sequencing templates, following standard methodologies and sequenced. Standard computational procedures, and publicly accessible databases were employed in analyzing the resulting sequences. There were some cases where the analysis revealed two, different cDNAs in the clone. This could be determined, since the primer sequence is present only at both ends of the cDNA. Thus, if the primer was found in the middle of the sequence, it indicated that the sequences on either side were from different cDNAs. The two sequences were treated as separate sequences in analyzing the results.

Of 413 cDNA sequences studied, 337 were not found in the public databases referred to, supra. Sixteen of these sequences had a partial match to known sequences, allowing a contig to be formed.

There were another 42 sequences which were similar, but not identical to, sequences in public databases, suggesting that these 42 sequences are related to the pre-existing material.

Twenty six of the sequences were completely contained within known, complete human sequences. This permitted generation of the empirical curve shown in Figure 2. Twenty two of the twenty six sequences were completely or partially within open reading frames of known genes.

Some of the sequences obtained showed partial homology to known genes, suggesting their function. Other sequences were found which showed no homology to known sequences.

EXAMPLE 7

This example shows the use of the invention as applied to breast cancer cells.

A sample of an infiltrative breast carcinoma with attached portions of normal tissues was operatively resected from a subject. The material was kept at -70°C until used. The sample was characterized, inter alia, by a large tumor mass and a very small amount of normal tissue.

Three x 20 micron-thick slices were taken across the tumor mass and any attached normal tissue was microdissected out to leave "pure" tumor tissue. One slice was treated to remove mRNA, as described, supra. Three cDNA libraries were prepared, using SEQ ID Nos: 8 & 9, as well as

5' - AGGAGTGACG GTTGATCAGT - 3' SEQ ID NO: 13

Reverse transcription was carried out as with the colon cancer sample, as described supra. Then, PCR amplification was carried out by combining 12.8uM of the same primer used in the reverse transcription 125uM of each dNTP, 1.5 mM $MgCl_2$, 1 unit of thermostable DNA polymerase, and buffer (50mM KCl, 10mM Tris-HCl, pH 9.0, and 0.1% Triton X-100), to a final volume of 20ul. Amplification was carried out by executing 1 cycle (denaturation at 94°C for 1 minute, annealing at 37°C for 2 minutes, and extension at 72°C, for 2 minutes), followed by 34 cycles at 94°C for 45 seconds, annealing at 55°C for 1 minute and extension at 72°C for 5 minutes. When analyzed for banding, as described supra, the samples revealed a complex pattern.

The products were eluted from their gels, cloned into pUC-18, and the plasmids were transformed into E. coli strain DH5α, all as described supra. Plasmids were subjected to minipreparation, using the known alkaline lysis method, and then about 150 of the molecules were sequenced. Of these, 69% were not found in any databank consulted, and appear to represent new sequences. A total of 22% was characterized by large quantities of repetitive elements and retroviral sequences. A total of 4% corresponded to known human sequences,

another 4% to ribosomal RNA and mitochondrial sequences, and 8% were redundant sequences.

EXAMPLE 8

An example of how a contig sequence can be built is described herein.

With reference to figure 3, the darker portion is a sequence obtained in accordance with the invention.

When the sequence was compared to sequences already accessible in databases, there was substantial overlap with a known sequence at the 3' end, and some overlap at the 5' end. This permitted construction of a 1,064 nucleotide long contig. The first sequence is a tentative human consensus sequence, as taught by Adams, et al., Nature 377: 3-17 (1995), while the third sequence is an EST obtained from human gall bladder cells, identified as human gall bladder EST 51121.

EXAMPLE 9

Following the experiments described supra, the sequences of the molecules identified therein were compared to sequences available to the public. One of these molecules, i.e., was found to be somewhat homologous to rat semaphorin Z (GENBANK Accession No: AB000776). The homology was found between nucleotides 780-897 of rat semaphorin Z and nucleotides 6-123 of SEQ ID NO: 1. The homology is to the midpoint of the semaphorin Z gene, in the coding region. Recent work by, e.g., Christensen, et al., Canc. Res 58:1238-44 (1998), and Martin-Satue, et al., J. Surg. Oncol 72:18-23 (1999), correlated expression of members of the semaphorin gene family with cancer. As such, work was undertaken to identify the full length sequence from which SEQ ID NO: 1 was derived.

To do this, RACE-PCR, in accordance with Frohman, et al., Meth. Enzymol 218: 340-356 (1993), incorporated by reference, was carried out on a pool of human mammary gland cDNA. To do this, one of the oligonucleotide primers:

cttggagtcatgttt cacgg

(SEQ ID NO: 2)

or

gggatgctcttcacagctact

(SEQ ID NO: 3)

which are 5' and 3' oriented gene specific primers were used, together with a primer that flanked either the 3' or 5' end of the cDNA molecule. More specifically, 5ngs of template were combined in a 25 μl PCR reaction, with 5 pmoles of each primer, 1.0 U rTth DNA polymerase, 800μM of each dNTP, 1.5mM $MgCl_2$, and 7.5μl of 3.3 X commercially available PCR buffer. Thirty-five cycles of PCR were carried out, one cycle being defined as 40 seconds at 94°C, 60 seconds as 60°C, and 120 seconds as 72°C. Before the thirty-five cycles, however, one touchdown program was run (one cycle of annealing as 72°C, two cycles as 68°C and 64°C, with identical denaturing and extension conditions). This resulted in an extension of the 5' end (i.e., nucleotides 321-887 of the final product, which is SEQ ID NO:6), but not the 3' end.

In view of this, a second set of experiments were carried out, using primers:

ccacgtggcatgcatggtcag

(SEQ ID NO: 4)

and

gccatgcagaccccgcgagc

(SEQ ID NO: 5)

These primers were designed in view of other known ESTs which showed high similarity to semaphorin Z, but in regions other than the regions to which SEQ ID NO: 1 is similar. RT-PCR was carried out, and this resulted in an 1801 nucleotide product. For the generation of this product, RT-PCR was carried out, as described supra, using an aliquot of 5ng of cDNA from a pool of mRNA from three isolated tumor breast tissues. In additional PCR experiments, SEQ ID NO: 4 or 5 was used with SEQ ID NO: 2 or 3, and these experiments generated sequence products approximately 0.8 and 1.0 Kb long.

The fragment products were separated by agarose gel electrophoresis, purified, cloned into pUC18, and sequenced using standard methods. The resulting1.8kb sequence is set forth as SEQ ID NO: 6 and represents the complete cDNA sequence . This sequence contains the entire 568 nucleotide fragment referred to supra (i.e., that obtained via RACE-PCR). The sequence exhibited 84% homology to murine semaphorin Z and semaphorin VIb which, together with presence of a sema domain, led to the conclusion that the molecule encodes a human semaphorin molecule.

The putative amino acid sequence of the protein encoded thereby is presented as SEQ ID NO: 7. A start codon is presumed to be at nucleotides 4-7 of SEQ ID NO: 6, and nucleotides 1555 represents the termination of the open reading frame. The open reading frame is 1551 base pairs long, spans 13 exons (exons 1-12 and exon 17), and ends with a 250 base pair 3' untranslated region.

Homology searches were then carried out, using standard methods. The sequence flanked by nucleotides 9 and 1394 of SEQ ID NO: 6 is 84% homologous to a sequence flanked by nucleotides 218 and 1606 of semaphorin Z, while the sequence flanked by nucleotides 1704 and 1801 of SEQ ID NO: 6 are 85% identical to nucleotides 3682 and 3779 of semaphorin Z. Further, the amino acid 1-169 of SEQ ID NO: 7 is 87% homologous to amino acids 1-166 of semaphorin Z. The region defined by nucleotides 321-566 of SEQ ID NO:6 showed no homology to any other molecule in publicly accessible data bases. Further, sequences defined by nucleotide fragments 72-566, 685-835, and 902-1410 were non-homologous with any ESTs associated with cancer.

EXAMPLE 10

This example describes further work on analysis of the semaphorin sequence described supra. The medium resolution Stanford G3 panel of DNA isolated from human/hamster hybrid cell lines, which is commercially available, was assayed, via PCR, using, as primers, SEQ ID NO: 3 and:

aggtagttaa actccatcgc aatc

(SEQ ID NO: 14). These primers were estimated to amplify a fragment of the semaphorin sequence described supra which was about 0.8 kb long, and contained an intronic sequence. The analysis indicated that the semaphorin R/6B locus was linked to the STS SHGC-1476 (lod_score = 5.6), located on chromosome 19, (distance of 36 cRs). The marker is not ordered on the map, but is linked to SHGC-3305 (lod_score = 3.2; distance of 60 cRs).

EXAMPLE 11

A BLAST search was then carried out, using the human sequence described herein, and the rat semaphorin Z sequence, against the HTGS database. This search indicated that there was a possible, alternatively spliced variant, with additional exons at the 3' end.

In order to evaluate this possibility, RT-PCR was carried out, using primers designed to amplify the potential, 3'- coding region that is absent from the initial sequence.

SEQ ID NO: 3 and

gaggagtttg agacctaccg gc

(SEQ ID NO: 15) were used. The results of the RT-PCR confirmed that there was, in fact, an alternate, longer human semaphorin sequence, the nucleotide sequence of which is set forth at SEQ ID NO: 16. The sequence contains an open reading frame 2061 base pairs long, which contains 4 additional exons, plus a cryptic acceptor site in the middle of the last exon, i.e., exon 17, in the 3' region of the mRNA sequence. This permits the RNA processing and posterior translation of the middle portion of the final exon. In contrast to the nucleotide sequence of SEQ ID NO: 6, the isoform encoded by this nucleotide sequence (687 amino acids), may contain a transmembrane domain, and a short cytoplasmic domain. The amino acid sequence is provided as SEQ ID NO: 17. Analysis using the "Pfam 5.2" database, as found at http://pfam.wustl.edu/, incorporated by reference, verifies this.

EXAMPLE 12

Previously, it had been observed that other semaphorins were involved in metastatic processes. See Christensen, et al., Canc. Res. 58(6): 1238-1244 (1998); Martin – Satué, et al, Surg. Oncol 72(1): 18-23 (1999); Eckhardt, et al., Mol. Cell Neurosci 9 (5/6): 409-419 (1997). To study whether this was the case with the newly identified sequences, their expression in human glioblastoma cell lines, regulated by antitumor agents, was tested. The lines T98G and A172, which display in vitro invasive activity, were used.

Glucocorticoid hormones or all-trans-retinoic acid (ATRA) were used to treat these cell lines for long periods of time, i.e., 24, 48 and 72 hours. (Glucocorticoid hormones are widely used as anti-inflammatory agents, and anti-tumoral agents, and are the only chemotherapeutic agents available for gliomas and glioblastomas). Retinoids are known to inhibit proliferation and migration of primary cultures of human multiform glioblastoma strongly, supporting clinical trials for their use. See Bouterfa, et al, Neurosurgery 46(2):419-430/2000)). Total RNA (10 ug) samples were isolated from cells that were untreated received glucocorticord treatement, or that received 10^{-5}M ATRA for the listed periods of time. Acid ribosomal phosphoprotein PO was used as an internal control for RNA loading.

The results indicated that glucocorticoid hormones did not regulate the expression of the sequence; however, the all-trans-retinoic acid did so, in a time dependent manner. Specifically, the expression was inhibited by 2.5 fold after 24 hours, and 7 fold after 72 hours in T98G, and in A172, expression was inhibited by 2.5 and 10 fold, respectively, after 48 and 72 hours.

In the case of T98G, a 3.8 kb band was observed which appears to represent SEQ ID NO: 6, and is downregulated by ATRA.

These results suggest that downregulation of the human semaphorin gene described herein underlies the anti-tumor action of all-trans-retinoic acid in human glioblastoma cells. This suggests a possible role for semaphorin gene product in tumor progression, which is consistent with prior reports on other members of the semaphorin family, including M-SemaH and semaphorin E, in the progression of murine mammary and human metastatic lung adenocarcinomas. See Christensen, et al., supra. Martin-Satué, et al., supra.

EXAMPLE 13

The expression of human semaphorin in various normal tissues was also tested via standard Northern blotting. A commercially available, human 12 lane multiple tissue Northern blot was used, where poly $(A)^+$ RNA was isolated, and hybridized to relevant human semaphorin sequences as described herein. GAPDH was used as an internal control.

The results indicated that there was very strong expression in brain (a 4.5kb band), weak expression in heart, spleen, placenta and lung with no expression in skeletal muscle, colon, thymus, kidney, liver, or small intestine tissue, or leukocytes.

The foregoing examples show that, using the ORESTES methodology described herein, an isolated nucleic acid molecule has been discovered which encodes for a molecule referred to herein as human semaphorin R/6B. "Human semaphorin R6B" as used herein refers to a protein encoded by, if nucleotides 4-1555 of SEQ ID NO. 6, where nucleotides 4-7 constitute a start codon, and nucleotides 1553-1555 a termination signal, as well as any protein which is encoded by the nucleic acid molecules of the invention, or is equivalent thereto such as proteins encoded by SEQ ID NO: 16. Also a part of the inventions are isolated nucleic acid molecules which encode this protein, such as the nucleotides which make up the 2061 base pair ORF of SEQ. ID NO:6, as well as nucleic acid molecules which comprise the

nucleotide sequence set forth in SEQ ID NO: 1 or 6. Also a part of the invention are isolated nucleic acid molecules which comprise the nucleotide sequence defined by nucleotides 321-566 of SEQ ID NO: 6. More particularly, those nucleic acid molecules which comprise nucleotides 72-566 of SEQ ID NO:6 are part of the invention. Also a part of the invention are nucleic acid molecules that comprise nucleotides 72-835 or nucleotides 72-1410 of SEQ ID NO:6. Expression vectors and recombinant cells which comprise these nucleic acid molecules are also a part of the invention. "Expression vector" as used herein, refers to any vector wherein the nucleic acid molecule is operably linked to a promoter. Recombinant cells in accordance with the invention are preferably eukaryotic cells, and may comprise the expression vector. With respect to the nucleic acid molecules, cDNA is preferred, but genomic DNA is also a part of the invention.

The proteins which are a part of this invention may be admixed with, e.g. , pharmaceutically acceptable adjuvants, such as those which are well known to the skilled artisan. Such compositions can be used, e.g., but not exclusively, to produce antibodies, such as monoclonal antibodies. These, as well as hybridomas producing them, are also a part of the invention. These proteins include e.g., those having amino acid sequences as set forth at SEQ ID NO: 7 or SEQ ID NO:17, as well as those proteins homologous thereto.

Expression of the nucleic acid molecules and proteins of the invention has been correlated to cancer, breast cancer in particular. Hence, yet another aspect of the invention is a diagnostic method for determining the possible presence of cancer, breast cancer or glioblastoma in particular, by determining expression or presence of one or both of the nucleic acid molecules and the proteins of the invention. One can carry out these assays via, e.g., DNA hybridization assays using SEQ ID NO: 2, 3, 4, 5, 14 or 15 antibody assays, and so forth.

Other aspects of the invention will be clear to the skilled artisan and need not be set forth herein.

The terms and expressions which have been employed are used as terms of description and not of limitation, and there is no intention in the use of such terms and expressions of excluding any equivalents of the features shown and described or portions thereof, it being recognized that various modifications are possible within the scope of the invention.

WE CLAIM

1. An isolated nucleic acid molecule which comprises the nucleotide sequence set forth at SEQ ID NO:1.

2. The isolated nucleic acid molecule of claim 1, comprising nucleotides 4-1555 of SEQ ID NO: 6.

3. The isolated nucleic acid molecule of claim 2, comprising SEQ ID NO: 16.

4. The isolated nucleic acid molecule of claim 1, wherein said molecule is cDNA.

5. The isolated nucleic acid molecule of claim 1, wherein said molecule is genomic DNA.

6. Expression vector comprising the isolated nucleic acid molecule of claim 1, operably linked to a promoter.

7. Recombinant cell comprising the isolated nucleic acid molecule of claim 1.

8. The recombinant cell of claim 1, wherein said cell is a eukaryotic cell.

9. Recombinant cell comprising the expression vector of claim 6.

10. The recombinant cell of claim 9, wherein said cell is a eukaryotic cell.

11. Isolated protein, the amino acid sequence of which comprises the amino acid sequence of SEQ ID NO: 7 or SEQ ID NO: 17.

12. Composition comprising the isolated protein of claim 11, and a pharmaceutically acceptable adjuvant.

13. Antibody which specifically binds to the isolated protein of claim 11.

14. The antibody of claim 13, wherein said antibody is a monoclonal antibody.

15. Hybridoma cell line which produces the monoclonal antibody of claim 14.

16. Method for determining possible presence of cancer in a sample comprising assay in a sample taken from a patient believed to have cancer for expression of the isolated nucleic acid molecule of claim 1 or a protein encoded by said isolated nucleic acid molecule, wherein expression of said nucleic acid molecule or said protein is indicative of possible presence of cancer.

17. The method of claim 16, wherein said cancer is breast cancer or glioblastoma.

18. The method of claim 16, comprising contacting said sample with a pair of oligonucleotide primers, each of which is from 17 to 50 nucleotides in length, each of which is complementary to the isolated nucleic acid molecule of SEQ ID NO: 1, 6 or 16.

19. The method of claim 15, comprising contacting said sample with the oligonucleotides, set forth in SEQ ID NOS: 2 and 4, or SEQ ID NOS: 3 and 5.

20. An isolated nucleic acid molecule which comprises nucleotides 321-566 of SEQ ID NO:6.

21. The isolated nucleic acid molecule of claim 20, comprising nucleotides 72-566 of SEQ ID NO:6.

22. The isolated nucleic acid molecule of claim 20, comprising nucleotides 72-835 of SEQ ID NO:6.

23. The isolated nucleic acid molecule of claim 20, comprising nucleotides 72-1410 of SEQ ID NO:6.

24. An isolated nucleic acid molecule which hybridizes to SEQ ID NO: 1, 6 or 16 and consists of at least 17 nucleotides and no more than 50 nucleotides.

25. The isolated nucleic acid molecule of claim 24, selected from the group consisting of SEQ ID NO:2, SEQ ID NO:3, SEQ ID NO:4, SEQ ID NO:5, SEQ ID NO: 14 or SEQ ID NO: 15.

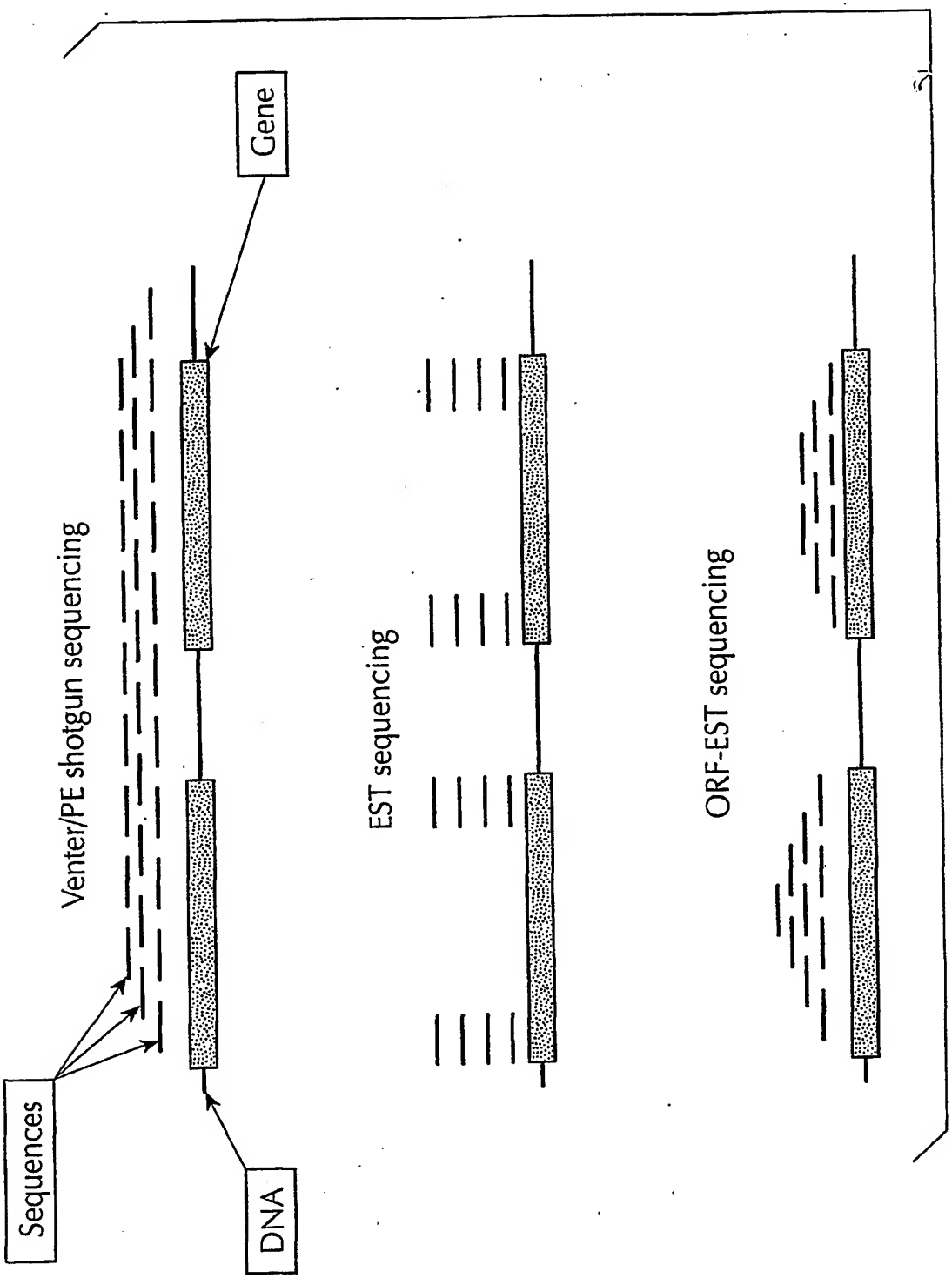

Gene
Venter/PE shotgun sequencing
EST sequencing
ORF-EST sequencing
Sequences
DNA

FIG. 1

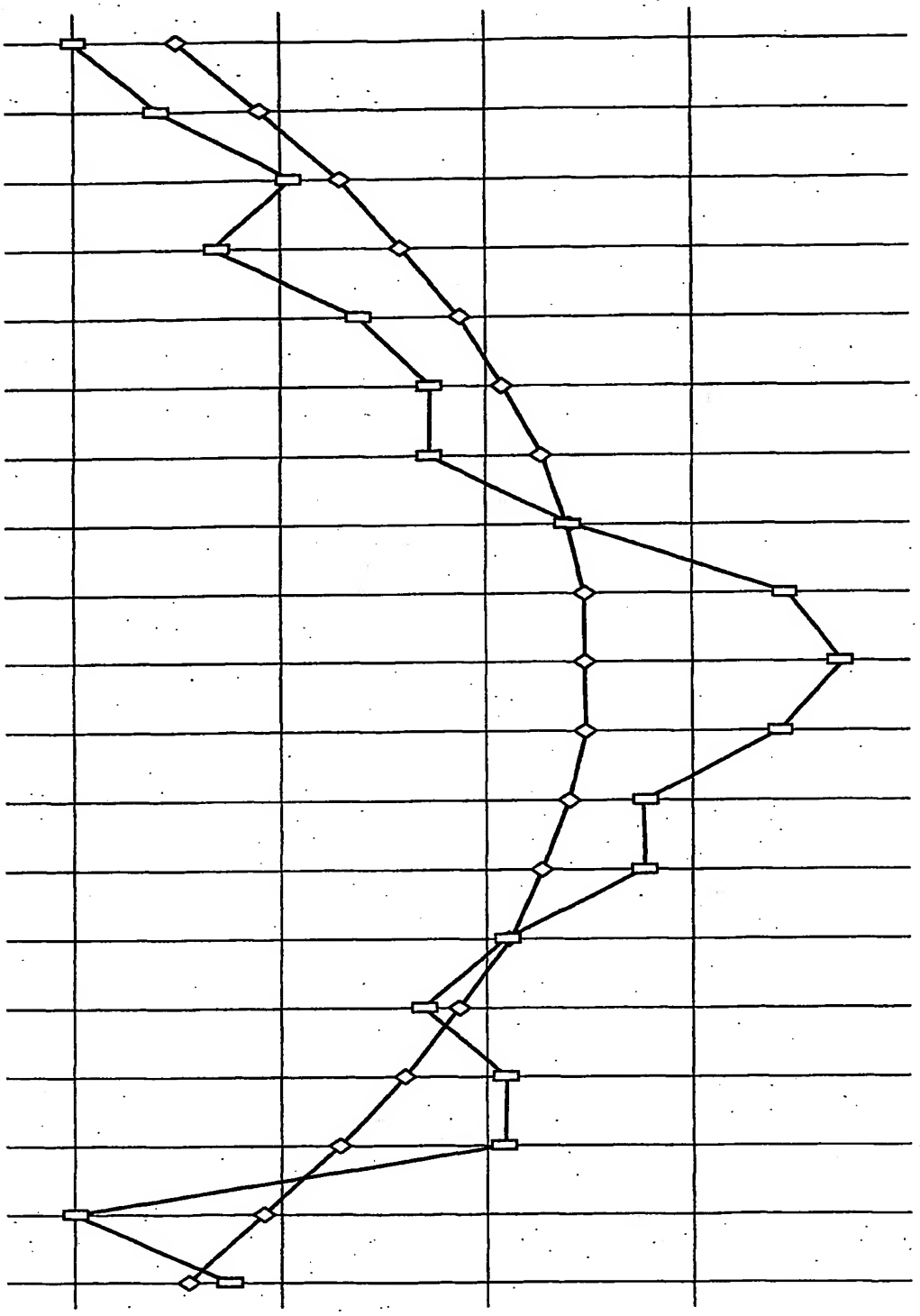


FIG. 2

attttaaataaaattgcccatcctcattcagctcttagaacaaaagcaaaaaaccctgt
aaatcaggagatataagcacatctgcacccagaataggcccatatgatagggcaaccctg
agcttaaacaatgacatcttcaagggtagaactaatctgaaaccccccttccagcctctgg
aagacactggcctgcatcagttagagtcagagcaagtgtcacttcacagggaaaagaagg
attatatagacttcctatccctagagtttataaatgtcaactatataaaaaaagctcaaa
acagtgttaaaggaatgaacagtaggaattttaataggctgtccaaagaagccaggtctg
ctgtgggcaagtatagcctaaccctagtcttgt

agtatagcctaaccctagtcttgtaaaataagccagaaagggttacttgag
ccacctttaagctagtacctatatagtaggcaaaaagtacagaaatagatgcaatagtgt
ggtgagtctttgagcctacgagtcatgccaccagccataagttgacctatcacttgagaa
cctcctcagcaaagatgccagaaaacattcaatcaagttggcaaatgacacagggagctg
gccctctgaccatcttcnctggcaaacctggactggaagggccatttgcagcactgtcct
ggagctaatacactgtttcactgcctctgccatataatgatgccagcactagccagctgg
tgggtatttggaggaattcctgcatgaggattgcccaataaggggcaggtacacatacct

cacacacct

ggcaaagtgatgatgatgtgaattgaaaccagtgaggggattgagtcaaaacttggatct
ggcaaagtgatgatgatgtgaattgtttccagtgaggggattgagtcaaaacttggntct

caggtacctcaatttttcccccaatttctggctactactaaaagccagaaagaacagaac
caggtacctcanttttttcccccaatttctggctactactaaaagccagaaagaacagaac

agtggcctcaggagatctgag-tttgaatcca
agtggcctcaggagatctgaggtttgaatccttgctctctaggatgcaggtggcttgaagc
agnatgccacaactncaagttgattagaactgccctttnttcccagggnttgacataggna
ttaagtcaaaaattncatggaaacccagtgggtaaaaaagcct

FIG. 3

<110> Simpson, Andrew George
Correa, Ricardo
deSouza, Sandro

<120> Isolated nucleic acid molecule encoding human semaphorin R and uses thereof

<130> LUD 5642.1 PCT

<140>
<141>

<150> US 09/483,618
<151> 2000-01-14

<150> US 09/196,716
<151> 1998-11-12

<150> US 09/406,117
<151> 1999-09-27

<160> 17

<210> 1
<211> 280
<212> DNA
<213> Homo sapiens
<220>
<400> 1
catacgggat gctcttcaca gctactgtta ccgacttcct agccattgat gctgtcatct 60
accgcagcct cggggacagg cccaccctgc gcaccgtgaa acatgactcc aagtggttca 120
aaggtgagac caggaagggc agtgggccca gacctggcag ggcccagaac ctgacattca 180
tcaaatctcc ccaccagaat aagagtccca gaggagcttg gtggagagat ttccaccaca 240
tctttggaat ctgttggcct ttttgaagac tagaaagcta 280

<210> 2
<211> 20
<212> DNA
<213> Homo sapiens
<220>
<400> 2
cttggagtca tgtttcacgg 20

<210> 3
<211> 21
<212> DNA
<213> Homo sapiens
<220>
<400> 3
gggatgctct tcacagctac t 21

<210> 4
<211> 21
<212> DNA
<213> Homo sapiens

<220>
<400> 4
ccacgtggca tgcatggtca g 21
<210> 5
<211> 20
<212> DNA
<213> Homo sapiens
<220>
<400> 5
gccatgcaga ccccgcgagc 20

<210> 6
<211> 1801
<212> DNA
<213> Homo sapiens
<220>
<400> 6
gccatgcaga ccccgcgagc gtcccctccc cgcccggccc tcctgcttct gctgctgcta 60
ctgggggggcg cccacggcct ctttcctgag gagccgccgc cgcttagcgt ggcccccagg 120
gactacctga accactatcc cgtgtttgtg ggcagcgggc ccggacgcct gacccccgca 180
gaaggtgctg acgacctcaa catccagcga gtcctgcggg tcaacaggac gctgttcatt 240
ggggacaggg acaacctcta ccgcgtagag ttggagcccc ccacgtccac ggagctgcgg 300
taccagagga agctgacctg gagatctaac cccagcgaca taaacgtgtg tcggatgaag 360
ggcaaacagg agggcgagtg tcgaaacttc gtaaaggtgc tgctccttcg ggacgagtcc 420
acgctctttg tgtgcggttc caacgccttc aacccggtgt gcgccaacta cagcatagac 480
accctgcagc ccgtcggaga caacatcagc ggtatggccc gctgcccgta cgaccccaag 540
cacgccaatg ttgccctctt ctctgacggg atgctcttca cagctactgt taccgacttc 600
ctagccattg atgctgtcat ctaccgcagc ctcggggaca ggcccaccct gcgcaccgtg 660
aaacatgact ccaagtggtt caaagagcct tactttgtcc atgcggtgga gtggggcagc 720
catgtctact tcttcttccg ggagattgcg atggagttta actacctgga gaaggtggtg 780
gtgtcccgcg tggcccgagt gtgcaagaac gacgtgggag gctccccccg cgtgctggag 840
aagcagtgga cgtccttcct gaaggcgcgg ctcaactgct ctgtacccgg agactcccat 900
ttctacttca acgtgctgca ggctgtcacg ggcgtggtca gcctcggggg ccggcccgtg 960
gtcctggccg ttttttccac gcccagcaac agcatccctg gctcggctgt ctgcgccttt 1020
gacctgacac aggtggcagc tgtgtttgaa ggccgcttcc gagagcagaa gtcccccgag 1080
tccatctgga cgccggtgcc ggaggatcag gtgcctcgac cccggcccgg gtgctgcgca 1140
gccccccggga tgcagtacaa tgcctccagc gccttgccgg atgacatcct caactttgtc 1200
aagacccacc ctctgatgga cgaagcggtg ccctcgctgg gccatgcgcc ctggatcctg 1260
cggaccctga tgaggcacca gctgactcga gtggctgtgg acgtgggagc cggcccctgg 1320
ggcaaccaga ccgttgtctt cctgggttct gaggcgggga cggtcctcaa gttcctcgtc 1380
cggcccaatg ccagcacctc agggacgtct gggcgtgtgt gtcaagtggg ccacgcgtgc 1440
agggtgtgtg tccacgagcg acgatcgtgg tggccccagc ggcctgggcg ttggctgagc 1500
cgacgctggg gcttccagaa ggcccggggg cctccgaggt gccggttagg agtttgaacc 1560
ccccccactc tgcagaggga agcggggaca atgccggggt ttcaggcagg agacacgagg 1620
agggcctgcc cggaagtcac atcggcagca gctgtctaaa gggcttgggg gcctgggggg 1680
cggcgaaggt gggtggggcc cctctgtaaa tacggcccca gggtggtgag agagtcccat 1740
gccacccgtc cccttgtgac ctccccccctc tgacctccag ctgaccatgc atgccacgtg 1800
g 1801

<210> 7
<211> 517
<212> PRT
<213> Homo sapiens
<220>
<400> 7
Met Gln Thr Pro Arg Ala Ser Pro Pro Arg Pro Ala Leu Leu Leu Leu

```
1               5                  10                  15
Leu Leu Leu Leu Gly Gly Ala His Gly Leu Phe Pro Glu Glu Pro Pro
            20                  25                  30
Pro Leu Ser Val Ala Pro Arg Asp Tyr Leu Asn His Tyr Pro Val Phe
        35                  40                  45
Val Gly Ser Gly Pro Gly Arg Leu Thr Pro Ala Glu Gly Ala Asp Asp
    50                  55                  60
Leu Asn Ile Gln Arg Val Leu Arg Val Asn Arg Thr Leu Phe Ile Gly
65                  70                  75                  80
Asp Arg Asp Asn Leu Tyr Arg Val Glu Leu Glu Pro Pro Thr Ser Thr
                85                  90                  95
Glu Leu Arg Tyr Gln Arg Lys Leu Thr Trp Arg Ser Asn Pro Ser Asp
            100                 105                 110
Ile Asn Val Cys Arg Met Lys Gly Lys Gln Glu Gly Glu Cys Arg Asn
        115                 120                 125
Phe Val Lys Val Leu Leu Leu Arg Asp Glu Ser Thr Leu Phe Val Cys
    130                 135                 140
Gly Ser Asn Ala Phe Asn Pro Val Cys Ala Asn Tyr Ser Ile Asp Thr
145                 150                 155                 160
Leu Gln Pro Val Gly Asp Asn Ile Ser Gly Met Ala Arg Cys Pro Tyr
                165                 170                 175
Asp Pro Lys His Ala Asn Val Ala Leu Phe Ser Asp Gly Met Leu Phe
            180                 185                 190
Thr Ala Thr Val Thr Asp Phe Leu Ala Ile Asp Ala Val Ile Tyr Arg
        195                 200                 205
Ser Leu Gly Asp Arg Pro Thr Leu Arg Thr Val Lys His Asp Ser Lys
    210                 215                 220
Trp Phe Lys Glu Pro Tyr Phe Val His Ala Val Glu Trp Gly Ser His
225                 230                 235                 240
Val Tyr Phe Phe Phe Arg Glu Ile Ala Met Glu Phe Asn Tyr Leu Glu
                245                 250                 255
Lys Val Val Val Ser Arg Val Ala Arg Val Cys Lys Asn Asp Val Gly
            260                 265                 270
Gly Ser Pro Arg Val Leu Glu Lys Gln Trp Thr Ser Phe Leu Lys Ala
        275                 280                 285
Arg Leu Asn Cys Ser Val Pro Gly Asp Ser His Phe Tyr Phe Asn Val
    290                 295                 300
Leu Gln Ala Val Thr Gly Val Val Ser Leu Gly Gly Arg Pro Val Val
305                 310                 315                 320
```

Leu Ala Val Phe Ser Thr Pro Ser Asn Ser Ile Pro Gly Ser Ala Val
325 330 335

Cys Ala Phe Asp Leu Thr Gln Val Ala Ala Val Phe Glu Gly Arg Phe
340 345 350

Arg Glu Gln Lys Ser Pro Glu Ser Ile Trp Thr Pro Val Pro Glu Asp
355 360 365

Gln Val Pro Arg Pro Arg Pro Gly Cys Cys Ala Ala Pro Gly Met Gln
370 375 380

Tyr Asn Ala Ser Ser Ala Leu Pro Asp Asp Ile Leu Asn Phe Val Lys
385 390 395 400

Thr His Pro Leu Met Asp Glu Ala Val Pro Ser Leu Gly His Ala Pro
405 410 415

Trp Ile Leu Arg Thr Leu Met Arg His Gln Leu Thr Arg Val Ala Val
420 425 430

Asp Val Gly Ala Gly Pro Trp Gly Asn Gln Thr Val Val Phe Leu Gly
435 440 445

Ser Glu Ala Gly Thr Val Leu Lys Phe Leu Val Arg Pro Asn Ala Ser
450 455 460

Thr Ser Gly Thr Ser Gly Arg Val Cys Gln Val Gly His Ala Cys Arg
465 470 475 480

Val Cys Val His Glu Arg Arg Ser Trp Trp Pro Gln Arg Pro Gly Arg
485 490 495

Trp Leu Ser Arg Arg Trp Gly Phe Gln Lys Ala Arg Gly Pro Pro Arg
500 505 510

Cys Arg Leu Gly Val
515

<210> 8
<211> 19
<212> DNA
<213> Homo sapiens
<220>
<400> 8
gaagctggta aacaaaagg 19

<210> 9
<211> 20
<212> DNA
<213> Homo sapiens
<220>
<400> 9
agctgcatga tgtgagcaag 20

<210> 10

<211> 20
<212> DNA
<213> Homo sapiens
<220>
<400> 10
cccgctcctc ctgagcaccc 20

<210> 11
<211> 17
<212> DNA
<213> Homo sapiens
<220>
<400> 11
gagtcgattt caggttg 17

<210> 12
<211> 17
<212> DNA
<213> Homo sapiens
<220>
<400> 12
tgcttaagtt cagcggg 17

<210> 13
<211> 20
<212> DNA
<213> Homo sapiens
<220>
<400> 13
aggagtgacg gttgatcagt 20

<210> 14
<211> 24
<212> DNA
<213> Homo sapiens
<220>
<400> 14
aggtagttaa actccatcgc aatc

<210> 15
<211> 22
<212> DNA
<213> Homo sapiens
<220>
<400> 15
gaggagtttg agacctaccg gc

<210> 16
<211> 2311
<212> DNA
<213> Homo sapiens
<220>
<400> 16
gccatgcaga ccccgcgagc gtcccctccc cgcccggccc tcctgcttct gctgctgcta 60

ctgggggcg cccacggcct ctttcctgag gagccgccgc cgcttagcgt ggcccccagg 120

gactacctga accactatcc cgtgtttgtg ggcagcgggc ccggacgcct gacccccgca 180
gaaggtgctg acgacctcaa catccagcga gtcctgcggg tcaacaggac gctgttcatt 240
ggggacaggg acaacctcta ccgcgtagag ttggagcccc ccacgtccac ggagctgcgg 300
taccagagga agctgacctg gagatctaac cccagcgaca taaacgtgtg tcggatgaag 360
ggcaaacagg agggcgagtg tcgaaacttc gtaaaggtgc tgctccttcg ggacgagtcc 420
acgctctttg tgtgcggttc caacgccttc aacccggtgt gcgccaacta cagcatagac 480
accctgcagc ccgtcggaga caacatcagc ggtatggccc gctgcccgta cgaccccaag 540
cacgccaatg ttgccctctt ctctgacggg atgctcttca cagctactgt taccgacttc 600
ctagccattg atgctgtcat ctaccgcagc ctcggggaca ggcccaccct gcgcaccgtg 660
aaacatgact ccaagtggtt caaagagcct tactttgtcc atgcggtgga gtggggcagc 720
catgtctact tcttcttccg ggagattgcg atggagttta actacctgga gaaggtggtg 780
gtgtcccgcg tggcccgagt gtgcaagaac gacgtgggag gctccccccg cgtgctggag 840
aagcagtgga cgtccttcct gaaggcgcgg ctcaactgct ctgtacccgg agactcccat 900
ttctacttca acgtgctgca ggctgtcacg ggcgtggtca gcctcggggg ccggcccgtg 960
gtcctggccg ttttttccac gcccagcaac agcatccctg gctcggctgt ctgcgccttt 1020
gacctgacac aggtggcagc tgtgtttgaa ggccgcttcc gagagcagaa gtcccccgag 1080
tccatctgga cgccggtgcc ggaggatcag gtgcctcgac cccggcccgg gtgctgcgca 1140
gcccccggga tgcagtacaa tgcctccagc gccttgccgg atgacatcct caactttgtc 1200
aagacccacc ctctgatgga cgaagcggtg ccctcgctgg gccatgcgcc ctggatcctg 1260
cggaccctga tgaggcacca gctgactcga gtggctgtgg acgtgggagc cggcccctgg 1320
ggcaaccaga ccgttgtctt cctgggttct gaggcgggga cggtcctcaa gttcctcgtc 1380
cggcccaatg ccagcacctc agggacgtct gggctcagtg tcttcctgga ggagtttgag 1440
acctaccggc cggacaggtg tgaacggccc ggcggtggcg agacagggca gcggctgctg 1500
gactcggggc tgctgagctt ggagctggac gcagcttcgg ggggcctgct ggctgccttc 1560
ccccgctgcg tggtccgagt gcctgtggct cgctgccagc agtactcggg gtgtatgaag 1620
aactgtatcg gcagtcagga cccctactgc gggtgggccc ccgacggctc ctgcatcttc 1680
ctcagcccgg gcaccagagc cgcctttgag caggacgtgt ccggggccag cacctcaggc 1740
ttaggggact gcacaggact cctgcgggcc agcctctccg aggaccgcgc ggggctggtg 1800
tcggtgaacc cgctggtaac gtcgtcggtg gcggccttcg tggtgagagc cgtggtgtcc 1860
ggcttcagcg tgggctggtt cgtgggcctc cgtgagcggc gggagctggc ccggcgcaag 1920

gacaaggagg ccatcctggc gcacggggcg ggcgaggcgg tgctgagcgt cagccgcctg 1980

ggcgagcgca gggcgcaggg tcccgggagc cgacgctggg gcttccagaa ggcccggggg 2040

tctccgaggt gccggttagg agtttgaacc ccccccactc tgcagaggga agcggggaca 2100

atgccggggt ttcaggcagg agacacgagg agggcctgcc cggaagtcac atcggcagca 2160

gctgtctaaa gggcttgggg gcctgggggg cggcgaaggt gggtggggcc cctctgtaaa 2220

tacggcccca gggtggtgag agagtcccat gccacccgtc cccttgtgac ctcccccctc 2280

tgacctccag ctgaccatgc atgccacgtg g 2311

<210> 17
<211> 687
<212> PRT
<213> Homo sapiens
<220>
<400> 17

<

Met Gln Thr Pro Arg Ala Ser Pro Pro Arg Pro Ala Leu Leu Leu Leu
1 5 10 15

Leu Leu Leu Leu Gly Gly Ala His Gly Leu Phe Pro Glu Glu Pro Pro
20 25 30

Pro Leu Ser Val Ala Pro Arg Asp Tyr Leu Asn His Tyr Pro Val Phe
35 40 45

Val Gly Ser Gly Pro Gly Arg Leu Thr Pro Ala Glu Gly Ala Asp Asp
50 55 60

Leu Asn Ile Gln Arg Val Leu Arg Val Asn Arg Thr Leu Phe Ile Gly
65 70 75 80

Asp Arg Asp Asn Leu Tyr Arg Val Glu Leu Glu Pro Pro Thr Ser Thr
85 90 95

Glu Leu Arg Tyr Gln Arg Lys Leu Thr Trp Arg Ser Asn Pro Ser Asp
100 105 110

Ile Asn Val Cys Arg Met Lys Gly Lys Gln Glu Gly Glu Cys Arg Asn
115 120 125

Phe Val Lys Val Leu Leu Leu Arg Asp Glu Ser Thr Leu Phe Val Cys
130 135 140

Gly Ser Asn Ala Phe Asn Pro Val Cys Ala Asn Tyr Ser Ile Asp Thr
145 150 155 160

Leu Gln Pro Val Gly Asp Asn Ile Ser Gly Met Ala Arg Cys Pro Tyr
165 170 175

```
Asp Pro Lys His Ala Asn Val Ala Leu Phe Ser Asp Gly Met Leu Phe
            180                 185                 190

Thr Ala Thr Val Thr Asp Phe Leu Ala Ile Asp Ala Val Ile Tyr Arg
        195                 200                 205

Ser Leu Gly Asp Arg Pro Thr Leu Arg Thr Val Lys His Asp Ser Lys
    210                 215                 220

Trp Phe Lys Glu Pro Tyr Phe Val His Ala Val Glu Trp Gly Ser His
225                 230                 235                 240

Val Tyr Phe Phe Phe Arg Glu Ile Ala Met Glu Phe Asn Tyr Leu Glu
                245                 250                 255

Lys Val Val Val Ser Arg Val Ala Arg Val Cys Lys Asn Asp Val Gly
            260                 265                 270

Gly Ser Pro Arg Val Leu Glu Lys Gln Trp Thr Ser Phe Leu Lys Ala
        275                 280                 285

Arg Leu Asn Cys Ser Val Pro Gly Asp Ser His Phe Tyr Phe Asn Val
    290                 295                 300

Leu Gln Ala Val Thr Gly Val Val Ser Leu Gly Gly Arg Pro Val Val
305                 310                 315                 320

Leu Ala Val Phe Ser Thr Pro Ser Asn Ser Ile Pro Gly Ser Ala Val
                325                 330                 335

Cys Ala Phe Asp Leu Thr Gln Val Ala Ala Val Phe Glu Gly Arg Phe
            340                 345                 350

Arg Glu Gln Lys Ser Pro Glu Ser Ile Trp Thr Pro Val Pro Glu Asp
        355                 360                 365

Gln Val Pro Arg Pro Arg Pro Gly Cys Cys Ala Ala Pro Gly Met Gln
    370                 375                 380

Tyr Asn Ala Ser Ser Ala Leu Pro Asp Asp Ile Leu Asn Phe Val Lys
385                 390                 395                 400

Thr His Pro Leu Met Asp Glu Ala Val Pro Ser Leu Gly His Ala Pro
                405                 410                 415

Trp Ile Leu Arg Thr Leu Met Arg His Gln Leu Thr Arg Val Ala Val
            420                 425                 430

Asp Val Gly Ala Gly Pro Trp Gly Asn Gln Thr Val Val Phe Leu Gly
        435                 440                 445

Ser Glu Ala Gly Thr Val Leu Lys Phe Leu Val Arg Pro Asn Ala Ser
    450                 455                 460
```

Thr Ser Gly Thr Ser Gly Leu Ser Val Phe Leu Glu Glu Phe Glu Thr
465 470 475 480

Tyr Arg Pro Asp Arg Cys Glu Arg Pro Gly Gly Gly Glu Thr Gly Gln
485 490 495

Arg Leu Leu Asp Ser Gly Leu Leu Ser Leu Glu Leu Asn Ala Ala
500 505 510

Ser Gly Gly Leu Leu Ala Ala Phe Pro Arg Cys Val Val Arg Val Phe
515 520 525

Val Ala Arg Cys Gln Gln Tyr Ser Cys Cys Met Lys Asn Cys Ile Gly
530 535 540

Ser Gln Asp Pro Tyr Cys Gly Trp Ala Pro Asp Gly Ser Cys Ile Phe
545 550 555 560

Leu Ser Pro Gly Thr Arg Ala Ala Phe Glu Gln Asp Val Ser Gly Ser
565 570 575

Thr Ser Gly Leu Gly Asp Cys Thr Gly Leu Leu Arg Ala Ser Leu Ser
580 585 590

Glu Asp Arg Ala Gly Leu Val Ser Val Asn Pro Leu Val Thr Ser Ser
595 600 605

Val Ala Ala Phe Val Val Arg Ala Val Val Ser Gly Phe Ser Val Gly
610 615 620

Trp Phe Val Gly Leu Arg Glu Arg Arg Glu Leu Ala Arg Arg Lys Asp
625 630 635 640

Lys Glu Ala Ile Leu Ala His Gly Ala Gly Glu Ala Val Leu Ser Val
645 650 655

Ser Arg Leu Gly Glu Arg Arg Ala Gln Gly Pro Gly Ser Arg Arg Trp
660 665 670

Gly Phe Gln Lys Ala Arg Gly Ser Pro Arg Cys Arg Leu Gly Val
675 680 685